Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 262 – 274

doi:10.3906/elk-1810-154

Lung cancer subtype differentiation from positron emission tomography images

Oğuzhan AYYILDIZ[1,*], Zafer AYDIN[2] , Bülent YILMAZ[1] , Seyhan KARAÇAVUŞ[3]
Kübra ŞENKAYA[4] , Semra İÇER[4] , Arzu TAŞDEMİR[5] , Eser KAYA[6]
[1]Department of Electrical and Electronics Engineering, School of Engineering,
Abdullah Gül University, Kayseri, Turkey
[2]Department of Computer Engineering, School of Engineering, Abdullah Gül University, Kayseri, Turkey
[3]Department of Nuclear Medicine, University of Health Science,
Kayseri Research and Training Hospital, Kayseri, Turkey
[4]Department of Biomedical Engineering, Faculty of Engineering, Erciyes University, Kayseri, Turkey
[5]Department of Pathology, Educational and Research Hospital, Kayseri, Turkey
[6]Department of Nuclear Medicine, School of Medicine, Acıbadem University, İstanbul, Turkey

Received: 22.10.2018 • **Accepted/Published Online:** 23.09.2019 • **Final Version:** 27.01.2020

Abstract: Lung cancer is one of the deadly cancer types, and almost 85% of lung cancers are nonsmall cell lung cancer (NSCLC). In the present study we investigated classification and feature selection methods for the differentiation of two subtypes of NSCLC, namely adenocarcinoma (ADC) and squamous cell carcinoma (SqCC). The major advances in understanding the effects of therapy agents suggest that future targeted therapies will be increasingly subtype specific. We obtained positron emission tomography (PET) images of 93 patients with NSCLC, 39 of which had ADC while the rest had SqCC. Random walk segmentation was applied to delineate three-dimensional tumor volume, and 39 texture features were extracted to grade the tumor subtypes. We examined 11 classifiers with two different feature selection methods and the effect of normalization on accuracy. The classifiers we used were the k-nearest-neighbor, logistic regression, support vector machine, Bayesian network, decision tree, radial basis function network, random forest, AdaBoostM1, and three stacking methods. To evaluate the prediction accuracy we performed a leave-one-out cross-validation experiment on the dataset. We also considered optimizing certain hyperparameters of these models by performing 10-fold cross-validation separately on each training set. We found that the stacking ensemble classifier, which combines a decision tree, AdaBoostM1, and logistic regression methods by a metalearner, was the most accurate method for detecting subtypes of NSCLC, and normalization of feature sets improved the accuracy of the classification method.

Key words: Machine learning, PET, lung cancer, texture analysis

1. Introduction

Almost 85% of lung cancers are nonsmall cell lung cancer (NSCLC) [1], and adenocarcinoma (ADC) and squamous cell carcinoma (SqCC) are the two major subtypes of NSCLC. ADC and SqCC correspond to about 40% and 25%–30% of lung cancers, respectively [2]. Until recently, therapeutic approaches to NSCLC were mainly guided by tumor stage, and there was no difference in treatment for ADC vs SqCC. The major advances in understanding the effects of cytotoxic and biological agents used in NSCLC therapy suggest that future targeted therapies will be increasingly subtype specific. Selection of patients for appropriate subtype specific therapies requires precise pathologic differentiation of ADC and SqCC [3]. The diagnosis of lung cancer is

*Correspondence: oguzhan.ayyildiz@agu.edu.tr

usually performed based on small biopsy (bronchoscopic, needle, or core biopsies) and cytology specimens. In most cases, the distinction of these two subtypes is achieved based on standard morphologic criteria by routine microscopy. However, distinction can be difficult in some poorly differentiated tumors, especially in small specimens. On the other hand, the characterization of the lesion by small biopsy might have a sampling error, which would not represent the actual biological behavior and the intratumoral heterogeneity. Positron emission tomography (PET) is a valuable functional imaging method. Its efficiency for patients with cancers of NSCLC to stage tumors, evaluate therapy response, define prognosis, and guide radiotherapy and surgery is proven. Recently, a concept called radiomics has become popular. The main hypothesis of radiomics is the following: medical images include more information than may be obtained by visual analysis [4]. Thanks to the increase in PET scanners' spatial resolution there has been a tendency among researchers towards using image processing tools/approaches for PET images. In this perspective, features extracted from PET images may help us to describe certain tumor properties in vivo at molecular level. Texture analysis is an approach that includes a set of pattern recognition and analysis methods. These methods are used to quantify the relationship between the pixels or voxels for better tumor characterization, monitoring and predicting of therapy response, and prognosis. Different combinations of textural features and automatic classification approaches have been utilized in different contexts such as predicting response to therapy and survival [5, 6] and tumor grade [7]. Computed tomography (CT) images have also been used for pulmonary nodule feature optimization [8], reproducibility and prognosis [9], and predicting survival [10]. In addition to medical imaging approaches like PET and CT, for lung cancer diagnoses automated quantitative analysis of histopathology images has been investigated [11, 12]. Machine learning studies the construction of algorithms that can learn from and make predictions on data to make intelligent decisions based on their recognition of complex patterns. The machine learning methods are used in oncology in different applications such as cancer prognosis and prediction [13], survival analysis [14], drug response [15], and gene expression [16]. The focus of the present study is medical image analysis and computer aided diagnosis. This is a classification problem in which the aim is to use PET images to determine whether a newly presented patient has a tumor subtype adenocarcinoma or squamous cell carcinoma; thus the oncological therapy may be guided accordingly. In a similar study [17] that aimed to cluster the subtypes using 24 textural features obtained from PET images, the researchers used linear discriminant analysis as the classification approach. In the present study we used 39 textural features that are frequently chosen by researchers to characterize tumor heterogeneity and analyzed the performances of different classification approaches that have not been utilized in tumor subtype discrimination in NSCLC.

2. Materials and methods

2.1. Patient population and PET/CT imaging

The present study includes 18F FDG PET/CT images of 93 patients with NSCLC. The imaging of patients was performed from March 2010 to April 2014 at Acıbadem Kayseri Hospital Nuclear Medicine Department, Kayseri, Turkey, using a PET/CT scanner (Siemens Biograph 6, HiRez). The Research Ethics Committee of Kayseri Research and Training Hospital (KRTH) approved this study. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. In this retrospective study the authors worked on the previously acquired images; thus they did not obtain informed consent from the participants. Out of the 93 patients 9 were female and 84 were male, with a mean age of 62.9 ± 4.5 (range: 39–84). The tumor subtypes of 39 patients were ADCs and of 54 patients were

SqCCs. The specimens were obtained using fine-needle or excisional biopsy and were assessed at the pathology department of KRTH in terms of tumor subtype.

2.2. Image processing and texture analysis

The methods that were used in the present study as a single shot are summarized in Figure 1. The details of the approaches shown in this figure are given in this and the following subsections of this section. For each patient PET and CT images were transferred to our computers. The study mainly focused on the PET images, especially slices with tumors (Figure 2). MATLAB (MathWorks, Natick, MA, USA) was used in the image processing steps of PET images in DICOM format. In the image processing part of the study, first the tumors were segmented in each slice, the image intensity values in the tumors in that slice were binned, and finally texture analysis approaches were applied to extract texture features from each three-dimensional tumor obtained by arranging two-dimensional slices in one stack. In the segmentation, a popular approach called random walk [18] was used to automatically distinguish the tumor from the background. Different segmentation methods like Otsu's, k-means, and active-contour approaches were also tested [19], and the best results were obtained using the random walk approach based on the comparison of the segmentation results and the manual drawings of the nuclear medicine expert in our team. The binning process corresponds to linear mapping of intensity values on the pixels of the segmented tumor region to be between 1 and 64. Various binning levels were tested, and 64 was found to be the optimal value, as it was previously [20] proved that levels more than 64 do not improve classification precision. In the final step, using four different texture analysis approaches from the binned regions with tumors 39 features were extracted. The approaches we used were the gray level cooccurrence matrix (GLCM, 8 features), gray level run length matrix (GLRLM, 13 features), gray level size zone matrix (GSZM, 13 features), and neighborhood gray tone difference matrix (NGTDM, 5 features). The details of these approaches can be found elsewhere [21]. The most common quantitative value derived from PET images that shows the uptake of radiotracer is the maximum standardized uptake value (SUVmax) in the tumor area, which is defined as the decay-corrected tumor activity concentration divided by injected activity per unit body weight, surface area, or lean body mass. In addition to the textural features we also included the SUVmax as the 40th feature, whose values ranged from 2.5 to 47.1 (15.5 ± 7.4).

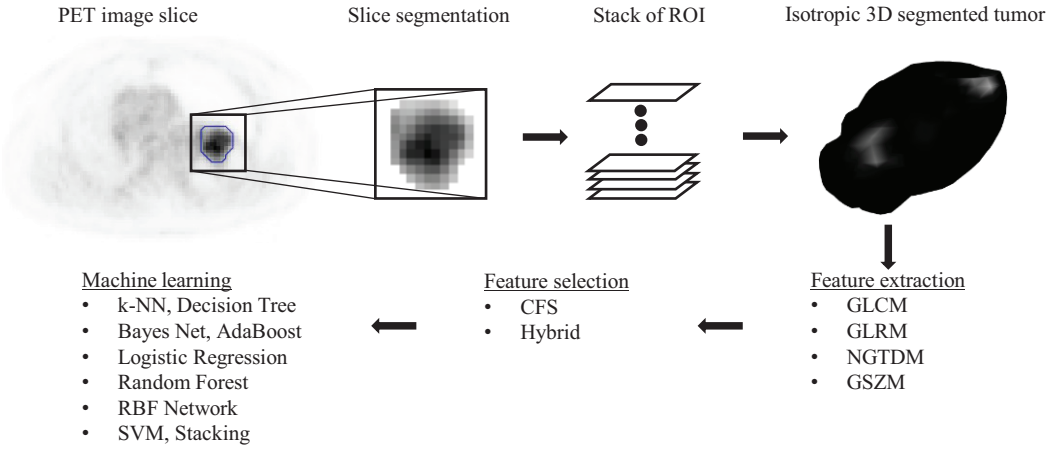

Figure 1. Summary of the approaches used in this study.

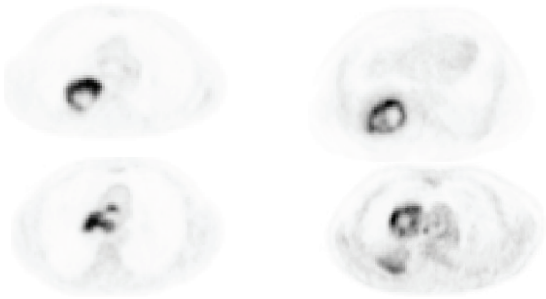

Figure 2. Sample images from raw data.

2.3. Data preprocessing and feature selection

We considered normalizing the texture features to the interval from 0 to 1. Feature selection methods are classified into three categories, namely filter, wrapper, and ensemble feature selection methods. The influence of the feature selection method on the performance of the classification method was examined before [16], and it was found that ensemble feature selection does not improve accuracy generally in breast cancer prognosis. To reduce the number of dimensions we implemented two feature selection methods in WEKA [22]: (1) CFS subset evaluator with BestFirst search strategy, (2) a hybrid strategy that first ranks features according to gain ratio, followed by a wrapper method that selects features using the k-NN classifier (with k parameter optimized by 10-fold cross-validation). In the wrapper approach we considered a sequential forward selection (SFS) strategy in which we start from an empty set and iteratively add the features to our feature set based on the initial feature ranking we obtained using gain ratio. Since it is recommended to have at least 1–10 samples per weight term for model training, the current size of the dataset will not be sufficient for an autoencoder model and would be prone to overfitting. Therefore, instead of projection-based methods such as autoencoders or PCA that map the input features to a new space, we considered feature selection methods, which allow selecting a subset of features, which also allows us to understand which features are important for correctly classifying the cancer subtypes.

2.4. Classification methods

In the present study, we implemented 11 different classifiers in WEKA software to differentiate the ADC and SqCC tumor subtypes: k-nearest-neighbor (k-NN), logistic regression, support vector machine (SVM), Bayesian network, decision tree, radial basis function (RBF) network, random forest, AdaBoostM1, and three stacking methods. We chose these classifiers for the same reason as Parmar et al. [14] did due to their popularity in the literature. To evaluate the prediction accuracy we performed a leave-one-out cross-validation (LOOCV) experiment on the dataset. We also considered optimizing certain hyperparameters of these models by performing 10-fold cross-validation separately on each training set.

2.4.1. k-nearest neighbor

A k-NN classifier first finds the k training samples that are closest to the test example and combines the class labels of these nearest neighbors by majority voting [23]. In our experiments, we employed the IBk method in WEKA to implement the k-NN classifier. We considered selecting the number of nearest neighbors (i.e. the k parameter) as 3 as well as optimizing this parameter by including the –X option in the command-line, choosing the maximum number of nearest neighbors as number of samples-2 and setting the number of cross-validation folds to 10.

2.4.2. Logistic regression

As a special case of generalized linear models, the logistic regression classifier computes a weighted linear combination of input features, which is passed through a nonlinear activation function (e.g., a sigmoid). In binary classification, the class labels are assigned by comparing the output variable to 0.5. The decision boundaries of a logistic regressor are linear hyperplanes [23]. We employed the logistic classifier in WEKA, which implements a multinomial logistic regression method with a ridge estimator and quasi-Newton optimization procedure.

2.4.3. Support vector machines

An SVM classifier aims to solve a quadratic optimization problem [24] by mapping the training samples to a higher dimensional space and finding a linear separating hyperplane with maximum margin [25]. We implemented two SVMs with a radial basis function (RBF) kernel using the LIBSVM package in WEKA. In the first version we set the C parameter to 1.0 and to 1/number of features, while in the second model we optimized these hyperparameters by performing a grid search choosing $\mathcal{C} \in (2^{-5}, 2^{-4}, , 2^{15})$ and $\mathcal{C} \in (2^{3}, 2^{2}, , 2^{-15})$. At the end of this procedure we selected the particular pair that gave the best cross-validation accuracy, trained the SVM classifier using these optima, and performed predictions on the test sample.

2.4.4. Decision tree

A decision tree classifier contains nodes and directed edges (i.e. branches) connecting nodes with no cycles allowed. Each internal node represents a test on a feature and each branch the outcome of the test, which can be true or false. For a given feature vector, the tests are applied starting from the top (root) node down to the leaf nodes, which represent a class label (i.e. final decision). Hence, each path from root to a leaf node is a classification rule. We employed the J48 algorithm in WEKA (a successor of C4.5) under default parameters [26], in which the confidence threshold for pruning is set to 0.25 and the minimum number of instances per leaf is set to 2.

2.4.5. Bayesian network

Let $X = [x_0, x_1, x_2, , x_d]$ be the set of variables, where $x_0 = y$ is the output class variable and $x_1, x_2, , x_d$ represent input features. A Bayesian network B over variables in X is a directed acyclic graph (DAG) and a set of probability tables $B_P = p(xpa(x))|x \in X$, where $pa(x)$ is the set of parents of x. The probability distribution for X can be computed as $P(X) = {}_(xX)p(xpa(x))$. The classification problem can be stated as inferring the class variable $y = x_0$ given the set of input features $x = [x_1, x_2, , x_d]$. In this context, a BayesNet classifier f x→y is a function that maps an input feature vector x to class type y. The classifier is learned from a dataset containing samples over (x, y) and the learning process includes deriving a Bayesian network structure and the mapping function f. The classification process selects the particular class type that maximizes the a posteriori distribution P(y | x). In the present study, we employed the BayesNet classifier in WEKA software, which first discretizes the continuous valued features by employing the filter called weka.filters.unsupervised.attribute.NumericToNominal. We selected the search algorithm for learning the network structure as K2, which is a hill climbing algorithm restricted by an order of the variables and the estimator as SimpleEstimator, which computes the conditional probability tables (CPTs) directly from the data for a given network structure [27].

2.4.6. Radial basis function (RBF) network

A radial basis function network first clusters data and then fits a basis function to each cluster. In the second stage, the basis function outputs are sent to a linear classifier to predict the class type [28]. We employed the RBFNetwork classifier in WEKA, which uses the k-means clustering algorithm and fits symmetric multivariate Gaussians to the data in each cluster. The output of Gaussians, which constitute the basis functions, are directed to a logistic regression classifier to predict the class type. All data are normalized to zero mean and unit variance (i.e. Z-score normalization). We implemented two versions of the RBFNetwork. The first one uses two clusters, which is equal to the number of class types, and the second optimizes the number of clusters by cross-validation considering the following values: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25.

2.4.7. Random forest

A random forest classifier is an ensemble technique that combines multiple decision trees by weighted majority voting. Each tree receives a small subset of input features constituted by random selection and is trained on a separate training set, which is generated by bootstrap sampling (also known as bagging) [29]. The random forest is also robust against outliers and is less prone to overfitting. We implemented two versions of the RandomForest classifier in WEKA. The first one uses 100 trees and the second one optimizes the number of trees by performing cross-validation on each training set and considering the following alternatives for this parameter: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100.

2.4.8. AdaBoost

A boosting ensemble combines multiple classifiers through weighted averaging of classifier outputs. Different from bagging, the base learner at a given iteration is constructed according to the classification behavior of the previous learner, concentrating more on the misclassified examples. To construct the training set of the current classifier, the probability of selecting misclassified examples is increased and a bootstrap sampling procedure is used [30]. Although boosting can be prone to overfitting, it typically improves the overall classification accuracy. We employed the AdaBoostM1 method in WEKA by selecting DecisionStump as the base learner and implemented two versions of this classifier. The first one selects the number of iterations as 10, which is the default value, and the second optimizes this parameter by performing cross-validation on each training set considering the following values: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100.

2.4.9. Stacking

A stacking ensemble combines different types of classifiers, which serve as base learners through a meta learner [31]. Typically the number of base learners is smaller than in bagging or boosting. In the present study, we implemented three stacking ensembles by combining different classifiers. The first ensemble combines decision tree (i.e. J48 in WEKA) with AdaBoostM1 (Stacking 1); the second combines decision tree, AdaboostM1, and logistic regression (Stacking 2); and the third combines decision tree, AdaboostM1, logistic regression, and BayesNet classifiers (Stacking 3). In each method we employed logistic regression as the meta learner and used 10 iterations for AdaBoostM1, which is the default setting in WEKA.

2.5. Accuracy measures

We used the following measures to evaluate the prediction accuracy of the classifiers: sensitivity (or recall), specificity, positive predictive value (PPV), negative predictive value (NPV), Matthew's correlation coefficient

(MCC), F-measure, overall accuracy, and area under ROC curve (AUC) [21]. These are computed as

$$Sensitivity = \frac{TP}{TP + FP}, \tag{1}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{2}$$

$$PPV = \frac{TP}{TP + FP}, \tag{3}$$

$$NPV = \frac{TN}{TN + FN}, \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{5}$$

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{6}$$

$$OverallAccuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

where TP is true positives, FP is false positives, TN is true negatives, and FN is false negatives. AUC measure is computed by first ranking the predictions with respect to the decision scores and then shifting the decision threshold to compute TP and FP rate values of the ROC curve. Each horizontal move (i.e. a false positive) generates a rectangular region in the ROC curve and the cumulative sum of these areas gives our AUC estimate.

3. Result and discussion

We performed a leave-one-out cross-validation experiment on the main dataset and obtained the accuracy measures shown in Tables 1 to 3. Table 1 compares different classifiers when no normalization is applied and the hybrid feature selection strategy is used. Table 2 demonstrates the accuracy of classifiers when the data are normalized and hybrid feature selection is employed. Table 3 includes the accuracy measures of the stacking ensemble for all combinations of the following conditions: data are not normalized, data are normalized, no feature selection is performed, CFS subset evaluator is employed, and hybrid feature selection method is employed. According to these results, we achieved the best results with the decision tree approach and stacking classifiers when data are normalized and the hybrid feature selection is used. Because decision tree was among the base learners in all stacking methods implemented, we can conclude that the stacking ensemble does not improve the accuracy of its base learners further. Based on the results presented in Tables 1 to 3, we can also observe that feature selection in general increases the classification accuracy compared to the condition where no feature selection is employed. Comparing the two feature selection methods some classifiers are more accurate when the first feature selection method is used while the rest give better results with the second strategy. Similar behavior is observed for data normalization conditions and there is no winner takes all condition. Furthermore, hyperparameter optimization improved the prediction accuracy of certain classifiers but not all of them. This could be related to constraints imposed by having a small number of samples. Table 4 shows the confusion matrix for the LOOCV experiment. It is evident that when the tumor subtype is SqCC the prediction is more

successful, but the identification is harder for the ADC subtype. This is the main reason for having low specificity as opposed to having high sensitivity since the number of SqCC samples is higher than the ADC samples and a classifier typically gives more weight to correctly estimating the majority class. A second reason for this behavior could be associated with the dataset, in which the class types overlap and separating positives from negatives becomes a challenging task. Figure 3 shows the histogram of the number of features selected on each training set of the leave-one-out cross-validation (a total of 93 feature subsets) when the hybrid feature selection is employed and no normalization is applied. According to this figure, most of the time approximately 20 features are selected out of 40. Similar behavior is observed when the same experiment is repeated on normalized data. Figures 4 and 5 show the relative importance of the features when hybrid feature selection is employed on not normalized and normalized data, respectively. The horizontal axis shows the features used in the present study and the vertical axis represent the number of times a feature is selected when feature selection is repeatedly applied to each training set of the leave-one-out cross-validation. Comparing these plots, the important features are similar for the two normalization conditions. Finally, when the decision tree classifier is trained on the normalized version of the dataset with 93 samples (without performing any feature selection), the tree diagram shown in Figure 4 is obtained, which performs a test on a single attribute named RLV (run-length variance, a parameter extracted from gray-level run-length matrix). This is also consistent with the relative importance rankings of the features in Figures 4 and 5. Since a decision tree classifier inherently performs feature selection and is pruned during training the resulting model is a feature-selected version of the original data. Furthermore, due to its simplicity, it is interpretable and can be applied in clinical settings directly on future data. The main reason for having low specificity as opposed to having high sensitivity is that the number of SqCC samples is higher than the number of ADC samples, and a classifier typically gives more weight to correctly estimating the majority class. A second reason for this behavior could be associated with our dataset, in which the class types overlap and discriminating positives from negatives becomes a challenging task. There is constant effort when using computer-aided diagnosis for classifying lung cancer subtypes. Different imaging techniques have been used so far, such as autofluorescence bronchoscopy images (ABIs) [32] and computer tomography images [33] besides the PET images. Each imaging modality has its limitations due to various imaging factors such as acquisition and reconstruction. Another point we should note is the preprocessing and segmentation challenges for a fair comparison between modalities. In a previous study [32] the authors compared the extraction of features from HSV and RGB channels, and used logistic regression for 34 patients based on ABIs to differentiate lung cancer subtypes. They showed that the transformation increased the classification accuracy for the ABIs. The PET/CT modality is the most common imaging technique in clinical oncology due to the availability of combined anatomical and metabolic information on the tumor. The major disadvantage of PET imaging is the spatial resolution when compared to CT images. However, PET gives a better representation of metabolic activity and tumor behavior or histopathological features. In another study [33] the authors applied two deep learning architectures, DBN and CNN, to compare their performances with the k-nearest neighbors and SVM classifiers, which employed the features extracted using SIFT and fractal approaches for the discrimination of lung nodules. However, the numbers of extracted features and classifiers were limited and tumor volume was not taken into account. In a previous study [17], the authors employed linear discriminant analysis (LDA) for the classification of lung cancer subtypes in 30 patients. They extracted the textural features from only one PET slice (one image), which contained the maximum standard uptake value (SUVmax). This approach is limited in terms of representing the whole tumor volume. In our study, we extracted three-dimensional features from the tumor volumes of 93 patients and investigated the use of 9 different classification approaches.

Table 1. Accuracy measures of classifiers when no normalization is applied and the hybrid feature selection method is employed.

Method	Sensitivity	Specificity	PPV	NPV	MCC	F-Measure	Overall	AUC
k-NN (k=3)	73.33	61.11	75.86	57.89	0.34	74.58	68.75	63.94
k-NN (k opt)	80.00	25.00	64.00	42.86	0.06	71.11	59.38	58.61
Decision tree (J48)	66.67	30.56	61.54	35.48	-0.03	64.00	53.12	46.39
Bayes net	83.33	36.11	68.49	56.52	0.22	75.19	65.62	53.70
AdaBoostM1 (iterations=10)	93.33	41.67	72.73	78.95	0.43	81.75	73.96	53.98
AdaBoostM1 (#iterations opt)	93.33	36.11	70.89	76.47	0.37	80.58	71.88	58.10
Logistic regression	75.00	47.22	70.31	53.12	0.23	72.58	64.58	65.37
Random forest (#trees=100)	73.33	38.89	66.67	46.67	0.13	69.84	60.42	59.95
Random forest (#trees opt)	66.67	47.22	67.80	45.95	0.14	67.23	59.38	58.47
RBF network (#clusters=15)	75.00	33.33	65.22	44.44	0.09	69.77	59.38	53.33
RBF network (#clusters opt)	66.67	44.44	66.67	44.44	0.11	66.67	58.33	52.08
SVM default	100.00	11.11	65.22	100.00	0.27	78.95	66.67	59.86
SVM opt	73.33	36.11	65.67	44.83	0.10	69.29	59.38	55.97
Stacking 1	90.00	33.33	69.23	66.67	0.29	78.26	68.75	50.83
Stacking 2	90.00	30.56	68.35	64.71	0.26	77.70	67.71	54.21
Stacking 3	86.67	33.33	68.42	60.00	0.24	76.47	66.67	49.77

Table 2. Accuracy measures of classifiers when data are normalized and the hybrid feature selection method is employed.

Method	Sensitivity	Specificity	PPV	NPV	MCC	F-Measure	Overall	AUC
k-NN (k=3)	68.33	47.22	68.33	47.22	0.16	68.33	60.42	56.85
k-NN (k opt)	86.67	22.22	65.00	50.00	0.12	74.29	62.50	55.74
Decision tree (J48)	95.00	44.44	74.03	84.21	0.48	83.21	76.04	42.22
Bayes net	88.33	38.89	70.67	66.67	0.32	78.52	69.79	52.31
AdaBoostM1 (#iterations=10)	91.67	38.89	71.43	73.68	0.37	80.29	71.88	49.49
AdaBoostM1 (#iterations opt)	90.00	38.89	71.05	70.00	0.34	79.41	70.83	50.79
Logistic regression	76.67	52.78	73.02	57.58	0.30	74.80	67.71	67.31
Random forest (#trees=100)	83.33	38.89	69.44	58.33	0.25	75.76	66.67	59.95
Random forest (#trees opt)	71.67	58.33	74.14	55.26	0.30	72.88	66.67	61.20
RBF network (#clusters=15)	78.33	36.11	67.14	50.00	0.16	72.31	62.50	68.75
RBF network (#clusters opt)	53.33	47.22	62.75	37.78	0.01	57.66	51.04	56.94
SVM default	100.00	0.00	62.50	0.00	0.00	76.92	62.50	53.52
SVM opt	78.33	36.11	67.14	50.00	0.16	72.31	62.50	59.26
Stacking 1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	67.18
Stacking 2	95.00	44.44	74.03	84.21	0.48	83.21	76.04	68.94
Stacking 3	95.00	44.44	74.03	84.21	0.48	83.21	76.04	62.27

4. Conclusions

In this work, we compared the accuracy of several machine learning approaches for discriminating the two cancer subtypes: adeno and squamous cell lung cancer. We also analyzed the effect of feature selection and data normalization. The most accurate method was the stacking ensemble classifier that combines a decision

Table 3. Accuracy of stacking methods with respect to normalization and feature selection. S1: First stacking method, S2: Second stacking method, S3: Third stacking method, FS0: No feature selection is performed, FS1: CFS subset evaluator is employed, FS2: Hybrid feature selection is employed, N0: No data normalization, N1: Features are normalized.

Method	Sensitivity	Specificity	PPV	NPV	MCC	F-Measure	Overall	AUC
S1 FS0 N0	90.00	30.56	68.35	64.71	0.26	77.70	67.71	48.80
S1 FS1 N0	88.33	36.11	69.74	65.00	0.29	77.94	68.75	69.49
S1 FS2 N0	90.00	33.33	69.23	66.67	0.29	78.26	68.75	50.83
S1 FS0 N1	95.00	36.11	71.25	81.25	0.40	81.43	72.92	61.20
S1 FS1 N1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	65.23
S1 FS2 N1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	67.18
S2 FS0 N0	83.33	27.78	74.03	50.00	0.13	73.53	62.50	49.26
S2 FS1 N0	88.33	36.11	69.74	65.00	0.29	77.94	68.75	66.11
S2 FS2 N0	90.00	30.56	68.35	64.71	0.26	77.70	67.71	54.21
S2 FS0 N1	95.00	33.33	70.37	80.00	0.38	80.85	71.88	58.01
S2 FS1 N1	93.33	44.44	73.68	80.00	0.45	82.35	75.00	69.63
S2 FS2 N1	95.00	44.44	74.03	84.21	0.48	83.21	76.04	68.94
S3 FS0 N0	83.33	27.78	65.79	50.00	0.13	73.53	62.50	43.61
S3 FS1 N0	86.67	36.11	69.33	61.90	0.27	77.04	67.71	61.85
S3 FS2 N0	86.67	33.33	68.42	60.00	0.24	76.47	66.67	49.77
S3 FS0 N1	91.67	33.33	69.62	70.59	0.32	79.14	69.79	51.81

Table 4. Confusion matrix for Stacking 2 classifier when data are normalized and the hybrid feature selection is employed.

True \Pred	Pred = ADC	Pred = SqCC
True = ADC	18	21
True = SqCC	3	51

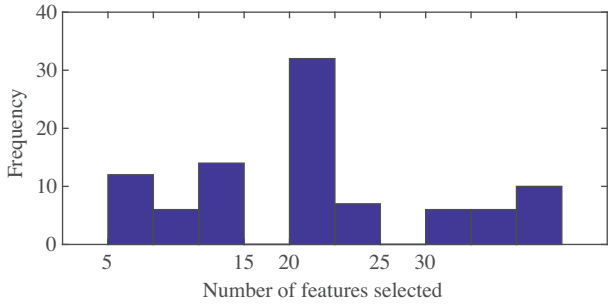

Figure 3. Histogram of the number of features selected on each training set of the leave-one-out cross-validation when no data normalization is performed.

tree, AdaBoostM1, and logistic regression methods by a meta learner. In future work, we are planning to test other feature selection methods in the machine learning literature and enlarge our dataset by including more subjects and new features. To improve the specificity of the classifiers, we are planning to apply threshold moving and adjust false positive as well as false negative rates according to the clinical expectations. All these

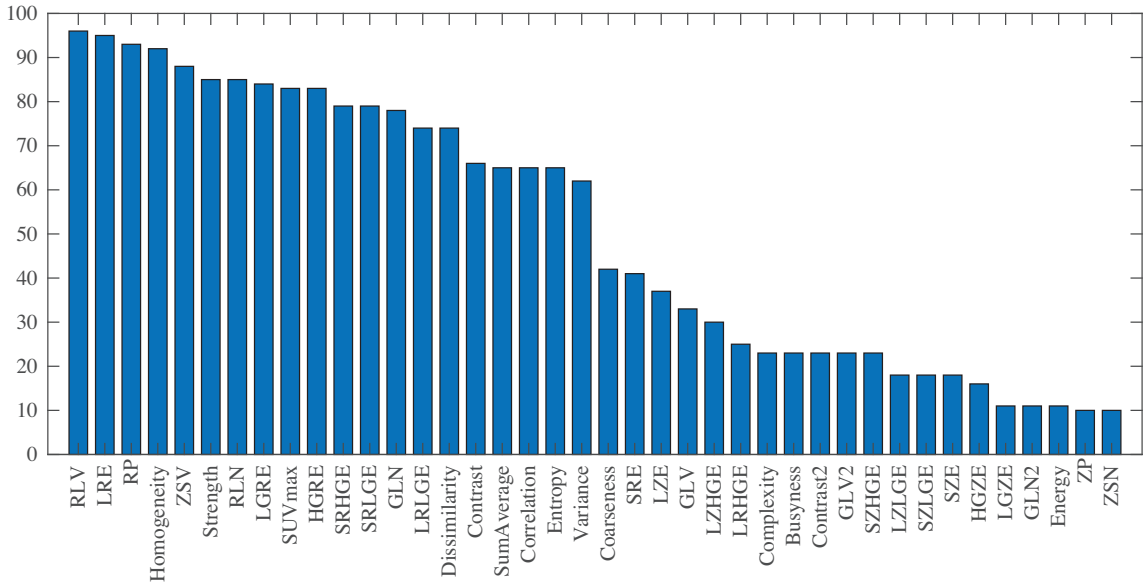

Figure 4. Selection frequencies of the features on training sets of the leave-one-out cross-validation when hybrid feature selection is employed and data are not normalized.

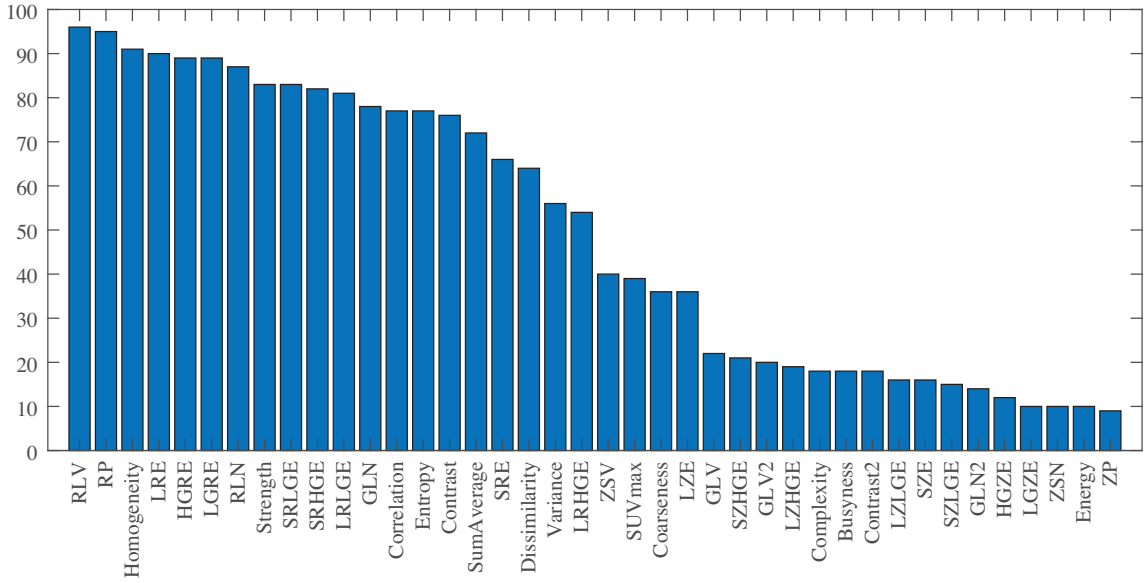

Figure 5. Selection frequencies of the features on training sets of the leave-one-out cross-validation when hybrid feature selection is employed and data are normalized.

efforts are expected to advance the detection of cancer subtypes, which is very important for future targeted therapies. In addition, in the literature, so far this kind of discrimination problem has not been handled in such a rigorous manner from feature selection to classification.

Acknowledgment

This study was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under Project No: 113E188.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Detterbeck FC, Boffa DJ, Tanoue LT. The new lung cancer staging system. Chest 2009; 136 (1): 260-271. doi: 10.1378/chest.08-0978

[2] Anagnostou VK, Dimou AT, Botsis T, Killiam EJ, Gustavson MD et al. Molecular classification of non-small cell lung cancer using a 4-protein quantitative assay. Cancer 2012; 118 (6): 1607-1618. doi:10.1002/cncr.26450.

[3] Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. Journal of Thoracic Oncology 2015; 10 (9): 1243-1260. doi: 10.1097/JTO.0000000000000630

[4] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA et al. Radiomics: the process and the challenges. Magnetic Resonance Imaging 2012; 30 (9): 1234-1248. doi: 10.1016/j.mri.2012.06.010

[5] Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. PloS One 2015; 10 (12): 1-16. doi: 10.1371/journal.pone.0145063

[6] Ypsilantis PP, Siddique M, Sohn HM, Davies A, Cook G et al. Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks. PloS One 2015; 10 (9): e0137036. doi: 10.1371/journal.pone.0137036

[7] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communication 2014; 5: 4006. doi: 10.1038/ncomms5006

[8] Samala R, Moreno W, You Y, Qian W. A novel approach to nodule feature optimization on thin section thoracic CT. Academic Radiology 2009; 16 (4): 418-427. doi: 10.1016/j.acra.2008.10.009

[9] Balagurunathan Y, Gu Y, Wang H, Kumar V, Grove O et al. Reproducibility and prognosis of quantitative features extracted from CT images. Translational Oncology 2014; 7 (1): 72-87. doi: 10.1593/tlo.13844

[10] Hawkins SH, Korecki JN, Balagurunathan Y, Gu Y, Kumar V et al. Predicting outcomes of nonsmall cell lung cancer using CT image features. IEEE Access 2014; 2: 1418-1426. doi: 10.1109/ACCESS.2014.2373335

[11] Aerts HJWL, Grossmann P, Tan Y, Oxnard GR, Rizvi N et al. Defining a radiomic response phenotype: a pilot study using targeted therapy in NSCLC. Science Report 2016; 6: 33860. doi: 10.1038/srep33860

[12] Yu KH, Zhang C, Berry GJ, Altman RB, Re C et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature Communication 2016; 7: 12474. doi: 10.1038/ncomms12474

[13] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 2015; 13: 8-17. doi: 10.1016/j.csbj.2014.11.005

[14] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. Scientific Report 2015; 5: 13087. doi: 10.1038/srep13087

[15] Menden MP, Iorio F, Garnett M, McDermott U, Benes CH et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PloS One 2013; 8 (4): e61318. doi: 10.1371/journal.pone.0061318

[16] Haury AC, Gestraud P, Vert JP. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS One 2011; 6 (12): 1-12. doi: 10.1371/journal.pone.0028210

[17] Ha S, Choi H, Cheon GJ, Kang KW, Chung JK et al. Autoclustering of non-small cell lung carcinoma subtypes on 18F-FDG PET using texture analysis: a preliminary result. Nuclear Medicine and Molecular Imaging 2014; 48 (4): 278-286. doi: 10.1007/s13139-014-0283-3

[18] Ju W, Xiang D, Zhang B, Wang L, Kopriva I et al. Random walk and graph cut for co-segmentation of lung tumor on PET-CT Images. IEEE Transactions on Image Processing 2015; 24 (12): 5854-5867. doi: 10.1109/TIP.2015.2488902

[19] Eset K, Icer K, Karacavus S, Yilmaz B, Kayaalti O et al. Comparison of lung tumor segmentation methods on PET images. In: Tiptekno 2015 Medical Technologies National Conference; Bodrum, Turkey; 2015. pp. 1-4 (in Turkish with an abstract in English).

[20] Leijenaar RTH, Nalbantov G, Carvalho S, Elmpt WJC, Troost EGC et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. Scientific Report 2015; 5: 11075. doi: 10.1038/srep11075

[21] Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Physics in Medicine and Biology 2015; 60 (14): 5471-5496. doi: 10.1088/0031-9155/60/14/5471

[22] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P et al. The WEKA data mining software: an update. SIGKDD Explorations 2009; 11 (1): 10-18. doi: 10.1145/1656274.1656278

[23] Bishop CM, Nasrabadi N. Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006. doi: 10.1117/1.2819119

[24] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20 (3): 273-297. doi: 10.1023/A:1022627411411

[25] Chang C, Lin C. LIBSVM: a library for support vector machines. ACM Transactions Intelligent System Technology 2013; 2: 1-39. doi: 10.1145/1961189.1961199

[26] Salzberg SL. Book Review: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning 1994; 16: 235-240. doi: 10.1023/A:1022645310020

[27] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning 1997; 29: 131-163. doi: 10.1023/A:1007465528199

[28] Lowe D. Multivariable functional interpolation and adaptive networks. Complex System 1988; 2: 321-355.

[29] Breiman L. Random forests. Machine Learning 1999; 45 (5): 1-35. doi: 10.1023/A:1010933404324

[30] Freund Y, Schapire RRE. Experiments with a new boosting algorithm. In: Thirteenth International Conference on ML; Bari, Italy; 1996. pp. 148-156.

[31] Wolpert DH. Stacked generalization. Neural Networks 1992; 5 (2): 241-259.

[32] Feng PH, Chen TT, Lin YT, Chiang SY, Lo CM. Classification of lung cancer subtypes based on autofluorescence bronchoscopic pattern recognition: a preliminary study. Computer Methods and Programs in Biomedicine 2018; 163: 33–38. doi: 10.1016/j.cmpb.2018.05.016

[33] Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets and Therapy 2015; 8: 2015-2022. doi: 10.2147/OTT.S80733